第十四章 測驗與評量

第一節 測驗的基本概念

一、測驗的發展

比奈	Binet	1905 年比西量表(Binet-Simon Scale)
馮德	Wundt	在萊比錫大學創立心理研究室為實驗心理學之父，奠定測驗，奠定測驗程序標準化基礎。
高爾登	Galton	致力於個別化差異研究，為評定量表、問卷方法、自由聯想技術的先驅。
卡泰爾	Cattell	美國測驗發展主要功臣，心理測驗一詞由他發明。
桑代克	Thorndike	1904 年出版第一本教育測驗的專書。
Gulliksen	1907/1987	古典測驗理論(classical test theory, CTT)之始
Cronbach, Gleser, Nanda, Rajaratnam,	1972	推論力理論(generalizability theory, GT)
Lord	1980	試題反應理論(item response theory, IRT)

二、測驗、測量、評定與評量

心理測驗(psychological test，了解人類心理的工具)與**心理測量**(psychological measurement，了解人類心理的手段)

測量	測驗	評定	評量
Measurement	Test	Assessment	Evaluation

三、測驗的種類

可分為**教育測驗**(educational test)與**心理測驗**(psychological test)兩種。

- **認知測驗、情意測驗與動作技能測驗**
  - 稱為能力測驗，包括智力測驗、性向測驗、成就測驗及創造思考測驗，最多是成就測驗(achievement test)。
  - **認知測驗**根據不同的教學用途分為
    - 用來測量各學科綜合成就水準的綜合成就測驗
    - 用來測量某單一學科成就水準的特殊成就測驗
    - 用來測量學生學習困難所在，以作為實施補救教學或學習輔導依據的診斷測驗。
  - **情意測驗**根據不同的教學用途分為
    - 專門用來測量個人對人事物等的看法、動機、興趣、價值觀、自我觀念等態度測驗(attitude test)，例如:民意調查。
    - 專門用來測量個人的人格特質、個性等的人格測驗(personality test)，知名的明尼蘇達多項人格特質量表(MMPI)即屬人格測驗的一種典型例子。
  - **動作技能**測驗根據不同的教學用途分為
    - 有關人的手、腳及腦等協調反應的測驗。較多以實作評量(Performance test)為主，輔以觀察、檢核表或評定量表等方式來進行。

- **標準化與非標準化測驗**
  - 根據不同編製過程來區分 standardized test & nonstandardized test

Standardized test	Nonstandardized test
只由專家根據測驗編序程序而編成的一種測驗。	指教師以非正式的方式，依自己教學的需要而自編的測驗，故又稱教師自編測驗(teacher-made test)或非正式測驗(informal test)。
大多數的智力測驗或是性向測驗	缺發標準化較為主觀，優點是可依班級做調整。

- **選擇反應測驗與結構反應測驗**

Selected-response test	Constructed-response test
選擇題、是非題、填充題、配合題、解釋性習題	簡答題、申論題、限制反應題、問答題
又稱客觀測驗(objective test)	又稱補充型式題測驗(supply-type items test)或論文測驗(essay test)

- **常模參照測驗**(Norm-referenced test, NRT)**與標準(校標)參照測驗**(Criterion-referenced test, CRT)
  源於葛拉瑟(Glasser, 1963)的分類方式
- **個人測驗**(individual test)與**團體測驗**(group test)
- **文字測驗**(verbal test)與**非文字測驗**(nonverbal test，又稱實作評量 performance test)
- **最大表現測驗與典型表現測驗**
  依據克朗巴哈(Cronbach, 1970)的分類模式

最大表現測驗(maximum performance test)	典型表現測驗(typical performance test)
目的要知道個人最佳反應獲最大成就，為確定個人能力表現。	正常情況下，表現的行為如何。沒有所謂的分數問題就只有好或是壞，二選一的問題。
智力測驗、性向測驗、成就測驗	人格測驗、興趣測驗、態度測驗、適應測驗
先天能力、實際能力、動機三項因素	

- **速度測驗與難度測驗**

速度測驗(speed test)	難度測驗(power test)
一定時間內可以做出的反應等，主要是測量速度的快慢。	主要是測量解決問題的能力。

- **客觀測驗**(objective test)與**非客觀測驗**(nonobjective test)

四．教學評量的功能

了解學生的起點行為	確定教學目標達到程度	做為改進教學的參考	評定學生的學習成果

- **落實輔導與諮商效能**
  輔導工作主要目的在增加個人的自我了解，消除或減少日常生活的困難與不適應行為，以充分實現其自我。

- **行政決定的功能**

選擇決定的功能	安置決定的功能
分類決定的功能	課程與教育計畫決定的功能

五．良好的成績

- **效度 0-1(low-high)**

  效度(validity)是指測驗的正確性，意旨測驗能夠測量到他所預測良知特質的程度，或是只測驗能達到其目的的程度。

- **信度 0-1(low-high)**

  信度(reliability)是指測驗的克靠信，包含測驗分數的穩定性與測驗內容的一致性。

- **常模**

  測驗結果要加以解釋，測驗的使用才有意義，就是要借助於常模才能達成任務。

- **實用性(practicality)**

經濟(易於取得所需的測驗)	容易實施	容易計分、解釋和應用

第二節 測驗的功能

一、信度

- **信度的意義**
  - 從測量的一致性來看，信度及指經由多次赴本測驗測量所得的結果的一致性(consistency)或穩定性(stability)。
  - 從測驗的誤差來看，信度是在估計測量誤差有多少，以反印出真實輛數(true measure)程度的一種指標
- **信度的原理**
  - **測驗有兩種結果**
    - 無關變異量(irrelevant variance) & 相關的變異量(valid or relevant variance)
  - **測驗所得的分數稱為實得分數**(observed score)

    Observed score= True score(universe) + Error score
  - **誤差種類**
    - Systematic error--- Random error
    - Unsystematic error--- Biased error(偏誤)
  - **信度係數**

    介於 0.00-1.00 越小代表性度越高，也代表誤差越小。
- **信度的類型**
  - 分析受試者本身的變異(intra-individual variability) 測量標準誤(standard error of measurement)
  - 分析受試者間的變異(inter-individual variability) 相關係數(correlation coefficient)
  - **常模參照測驗的信度分析**
    - **再測信度**

      重測一次所得兩個測驗的相關係數為重測信度係數(test-retest reliability coefficient)，或簡稱為再測信度(test-retest reliability；穩定信度 coefficient of stability)，時間越久信度就會降低。
    - **複本信度**
      - 根據雙向細目表建立題庫，製成兩份測驗所得的相關係數為複本信度(parallel-forms reliability)。
      - 一種方式是連續做測驗，數值稱為等值係數(coefficient od equivalence)
      - 一種方式是隔段時間再施測，數值稱為穩定且等值係數(coefficient of stability and equivalence)
    - **內部一致性信度**(internal consistency reliability)
      - **折半方法**(split-half correlation)

        單獨一次測驗結果，隨機方式分成兩半，再求其相關係數得知(Spearman-Brown Formula)

- **庫李方法**(K-R Method)二元論
  依據學生對所有試題的反應後，分析試題間的一致性，以確定測驗中的試題是否都測量到相同特質或相同單一能力的一種信度估計方法。誤差是屬於系統誤差為多，因為試題抽樣的同質性或是異質性。
- **α係數**(coefficient alpha)克朗巴哈(Cronbach)
  較屬於多元測驗的計分方式，李克特五點表。
  誤差與 K-R 一樣是抽樣誤差，尤其是異質性的誤差。
  α係數為所有信度的下限，故α係數頗高，真正的信度比它還高；或α係數低，無法得知信度的可靠性。

➢ **評分者信度**(scorer reliability)
- 上述使用在客觀測驗，此方法是用主觀測驗。
- Spearman 等級相關係數 & Pearson 積差相關法 & 評分者多 Kendall(1970)和諧係數(coefficient of concordance)

評分方式 \ 評分者人數	2 名	2 名以上	
名次法	Spearman 等級相關係數	Kendall(1970)和諧係數 (coefficient of concordance)	等級資料
分數法	Pearson 積差相關法	變異數分析(Hoyt 法)	等距資料

校標參照測驗的信度分析

➢ **百分比一致性指標**(percent agreement, P_A)
指標係指分析前後兩次分類決定結果是否一致的一種統計方法，並以百分比的和來表示。
P_A = 精熟人數/總人數 +未精熟人數/總人數 = 一致性的高低 (易高估)

➢ **K 係數**(Cohen's Kappa coefficient, *k*)
與前一個相同適用於類別與名義變相的信度分析。P_C 越高一致性越高。

影響信度的因素

試題題數(測驗長度)	樣本能力分配(團體分數的變異程度)	試題難易程度
測驗評分的客觀性	信度的估計方法 --- 因素分析&鑑別度分析	

信度係數的解釋與應用

- **理想的信度係數**(兩團體間比較 0.65---個人與他人間比較 0.85)
- **測量標準誤**

➢ **測量標準誤的意義**
測量標準誤(standard error of measurement, SEmeas.) & **分數標準誤**(standard error of a score)
有三種情況: (1) 0 誤差 (2)誤差是成常態分佈 (3)誤差分數的標準差就是測量表準誤

➢ **測量標準誤的應用**---信賴區間(confidence interval)

二、效度

- **效度的意義**

  在各種教育研究或教育歷程中，為了各種不同的目的，常需要運用測驗評估學生的心理特質或行為。效度就是指分數的正確性，是依分數的使用目的或結果來做解釋。

- **效度的原理**

  測驗分數必然與所要測量的潛在間具有某種程度的關係(及共同變異部分)存在。

- **效度的類型**

  - **內容效度**(context validity)

    - **涵意**

      是指測驗的試題能否適切地測到此測驗編制目的上所測的內容。

    - **驗證內容效度的方法**

      - **邏輯的分析方法**

        邀請專家學者來編製雙向細目表，仔細判斷每個測驗試題是否與教材內容所涵蓋的範圍及教學目標相符。

      - **實證的分析方法**

        除了專家判斷外，也有學者建議以內容信度取代之。

        ✓ 列出教材內容的主題及所預期的行為改變
        ✓ 判斷每一個主題的相對重要性
        ✓ 建立雙向細目表
        ✓ 依據雙向細目表編寫符合主題及行為改變的題目

    - **用途**

      適用教育測驗(成就測驗)；不適合用心理測驗(性向與人格測驗)

      1. 測驗是否涵蓋特定的知識或技能的代表性樣本
      2. 測驗成績是否不受無關因素的影響

  - **效標關聯效度**(criterion-related validity)

    - **涵義**

      指以實證分析方法研究測驗分數與外在效標間關聯性的一種指標，又稱「實證效度」或「統計效度」。(empirical or statistical vadility)

    - **種類**

      - 同時效度(concurrent validity) 測驗效標與外在效標同時得到。
      - 預測效度(predictive validity) 以測驗效標評估外來的外在效標的表現。

    - **效標特徵**

適切性	可靠性	客觀性	可用性
relevance	reliability	freedom from bias	availability

- **效標種類**

學業成就	特殊化的訓練成績	實際工作表現
Academic achievement	Performance in specialized training	Actual job-performance
智力測驗&成就測驗	針對某職業設計。 又有中間效標&終點效標	最令人滿意的效標示實際工作表現。

對照團體	相關人員的評定	精神病學的診斷	先前的有效測驗
Contrast groups	ratings	Psychiatric diagnosis	Previously available tests
音樂性向測驗就是跟普通省比較。	---	---	---

- **預期表**(expectancy table)---了解分數與效標間的關係。
- **使用的注意事項**

留意效標關聯效度產生變化的原因	留意外在效標的意義
留意適當的樣本大小(交叉驗證 cross validation)	留意效度推論的證據

建構效度

- **涵義**

  是指心理學或社會學上的一種理論構想或特質，它本身是觀察不到、也無法測量到，但卻被學術理論假餓是存在的，以便能夠用來解釋和預測個人或團體的行為表現。

- **驗證方法**
  - **內部一致性分析**(interval consistency)
    - ✓ **有關分析法:** 及計算每個試題的二元化得分和總分間的二系列相關或點二系列相關，凡相關係數經考驗後達到顯著水準者，即可保留，反之刪除淘汰。
    - ✓ **團體對照法:** 即依據學生的測驗總分高低，將學生分成高分組與低分組兩組，然後比較這兩組學生在每個試題上的答對百分比值。
  - **實驗研究**(experimental studies)

    比較實驗處理前後的分數差異，是建構效度的另一種資料來源。
  - **與外在效標的相關**(congruent validity)

    這種方法與效標關聯效度完全相同，但其目的不在證明他做理論或預期時的有效性，而是證明他與其他測量方法的測量工具的關係。
  - **因素分析**(factor analysis)---人格量表

    是一種多變量的統計方法，近年來與電腦硬體設備同步發展，線性結構關係模式(linear structure relationship model, LISREL)。
  - **多重性質**(multitrait-multimethod approach)
    - ✓ **收斂(幅合)效度**(convergent validity)

      即一份測驗分數要能夠和其他測量相同理論建構或潛在特質的測驗分數間具有高相關。
    - ✓ **區別(區辨)效度**(discriminant validity)

      即一份測驗分數也要能夠和其他測量不同理論建構或潛在特質的測驗分數具有高低相關。

- **表面效度**(face validity)
  - 只受測者、測驗結果的使用者即依般大眾對於該測驗的試題和形式等所做的主觀判斷，判斷此一測驗能否達成其所宣傳的目的。
  - **最常降低表面效度的原因**
    - 版面設計與印刷品質不良
    - 遣詞用字不能配合受測者的程度與背景
    - 指導手冊內容不夠完備，未能做充分的溝通

影響效度的因素

1.測驗編製過程是否得當	2.施策程序與情境是否良好	3. 外在效標品質
4.受試者的身心反應因素-反應心向(response set)	5.樣本能力分配的變異程度	篩選(preselection)
		放棄不用或重新編擬試題
		校正相關係數的萎縮

效度與信度的關係

- **信度是效度的必要條件而非充要條件**

信度高，效度不一定高。	信度低，效度一定低。
效度高，信度一定高。	效度低，信度不一定低。

- **效度與信度的變異數關係**

  信度 = 效度 + 獨特性

三．常模及測驗結果的解釋

理想的測量量尺具有絕對零點(absolute zero point)與相等的單位。

常模的意義與功用

- **常模的意義**
  - 常模(norm)式解釋測驗分數的依據。所謂常模，就是指特定參照團體(reference group)在測驗上鎖或的的平均分數。
  - 常模應該是視為學生的一種典型的表現(typical performance)或常態的表現(normal performance)。
  - 以試題取向的內容做為參照標準，又稱為內容參照(content reference)或效標參照(criterion-reference)
  - 以受試者團體的平均作為參照標準，稱為常模參照(norm-reference)，其轉換後的分數稱為轉換後分數或衍生分數(converted or derived scores)。
- **常模的功用**
  - 原始分數本身沒有多大意義，為了解釋其意義，測驗編制者都會提供各種常模表，便將不同類型的衍生分數(相對量數)來做解釋測驗分數，目的在使原本沒有絕對零點與相等的分數變為更具有意義。

新近性	代表性	適切性
社會各方面均在不斷改變中，由以教育為然。	測驗分數經常有兩種誤差，測量誤差、抽樣誤差。 **注意:** 抽樣的方法、常模樣本的分配與數量、常模團體的特性、測驗情境的標準化程度及受試的動機、測驗日期。	根據各種常模參照團體可以建立不同的常模類型。

常模的類型

全國性常模 (National norms)	地區性常模 (Local norms)	特殊團體常模 (Special group norms)	學校平均數常模 (School averages norms)

常模的建立方法

建立上述各類型的常模時，須將原始分數轉換成各種衍生分數。個人表現的不同發展層次做比較稱為發展性常模(developmental norm)。個人表現和他人做比較稱為組內(團體內)常模(within-group norm)。

發展性常模

年齡當量(年齡常模)	年級當量(年級常模)	順序量表	發展商數
Age equivalent	Grade equivalent	Ordinal equivalent	Development quotient

團體內(組內)常模

- 百分等級(PR, percentile rank)
- 標準分數(standard score)

直線轉換(linear transformation)	轉換	直線標準分數(linear standard score)
非直線轉換(linear transformation)	轉換	常態化標準分數(normalized standard scroe)

直線標準分數

Z 分數	T 分數	ETS 分數	AGCT 分數	離差智商
M = 0 SD = 1	M = 50 SD = 10	M = 500 SD = 100	M = 100 SD = 20	M = 100 SD = 15(W) SD = 16(S-B)
$Z = \frac{S-M}{SD}$	$T = 10Z + 50$	$ETS = 100Z + 500$ $CEEB = 100Z + 500$	$AGCT = 20Z + 100$	$DIQ = 15Z + 100(w)$ $DIQ = 16Z + 100(sb)$

常態化標準分數

標準九	C 量表分數	Sten 分數	常態化轉換	常態分配
分為九個組	分為 11 組	跟標準九一樣	同上	Normal distribution

測驗分數的解釋

解釋測驗分數的類型---高德門(Goldman, 1971)資料種類、資料處理方式、不同層次解釋

處理方式分為

- 機械的---統計的處理(statistical treatment)---常模對照表、預期表、回歸預測
- 非機械的---臨床診斷(clinical treatment)---歸納、演繹

解釋類型代表四種不同層次的解釋

敘述的解釋	溯因的解釋	預測的解釋	診斷的解釋
Descriptive interpretation	Genetic interpretation	Predictive interpretation	Evaluation interpretation

解釋測驗分數應注意的原則

1	解釋測驗者應了解測驗的性質與功能	2	測驗分數應為學生保密
3	解釋分數應參考其他有關資料	4	解釋分數應避免只給數字
5	對低分者的解釋應謹慎小心	6	解釋分數時應設法了解學生的感受
7	解釋分數只做建議，而勿做決定	8	應以一段可信範圍解釋

第三節 標準化測驗

一、涵義

- 經專家設計，其目的是為了能將一個學生的表現與其他同年級和同年齡的學生進行比較。
- 測驗的編制程序標準化，測驗的實施程序標準化。

二、標準化測驗的類型

	成就測驗	性向測驗
性質	評估一套以隻或控制情境下所得經驗的效應。	評估一套未知或非控制情境下所得經驗的效應。
能力	評估訓練課程的最後成果，及成就。	評估訓練課程所獲取的潛能。
效度	強調內容效度。	強調預測效度。

- **智力測驗與成就測驗**：智力是測一個人一般潛在能力；成就是測個人運用先天潛在能力。
- **性向測驗與成就測驗**：性向測驗測個人「學習」能力；成就測驗是測量個人「學習後」的結果。

成就測驗

- **涵義**：人類的能力可以解釋為人的一種認知或智能，但心理與教育測驗上常持以狹義的觀點，專指透過心理學技術所能測量到的心理能力(mental ability)。
- **種類**

依編制程序分	標準化成就測驗	
	教師自編成就測驗	
依測量科目分	綜合成就測驗(SAT, Stanford Achievement test)	
	單科成就測驗(TOFEL)	
依施測目的分	普通成就測驗	目的在了解學生依班學習狀況及學校成就，題目難度力求常態分佈，難度適中。
	診斷性成就測驗	(1)為低於平均數下的學生設計
		(2)可找出學生不會的原因
		(3)只診斷原因為討論其他因素
		(4)錯誤題目少，信度較低

智力測驗

- **性質**：如第五章。
- **種類**
  - **個別智力測驗**

比西量表	魏氏量表	考夫曼兒童智力測驗
Binet-Simon；Terman	Wechsler scale	KABC

  - **團體智力測驗**

文字測驗		非文字測驗	
普通分類測驗(AGCT)	加州心理成熟測驗(CTMM)	瑞士文非文字推理測驗(SPM)	修訂多尼(TONI)

測量智力的研究

智力分配	智力的穩定性	智力與學業成就	智力的文化公平測驗

性向測驗

- **性質**(一般能力 general aptitude；特殊能力 special aptitude)

  指一種測量受試者的學習潛能，工作推估未來接受學習或測驗可能成就的心理測驗。

- **種類**

特殊性向測驗(special aptitude test)		多因素性向測驗(multifactor aptitude test)	
心理動作測驗	創造性向測驗	教育導向性向測驗組合	職業導向性向測驗組合
Psychomotor tests	Convergent thinking Divergent thinking	區樣性向測驗(DAT) Differential aptitude test	通用性向測驗(GATB) General aptitude test battery

人格測驗

人格的測量問題

人格定義的問題	信度與效度的問題	反映新象的問題	偽裝答案的問題
人是獨特的統合體。	---	受試者使用一種特殊的方向回答題目。	情意測驗比認知測驗普遍且嚴重。

種類

- **自陳良表**(self-report inventories)

內容效度自陳量表	效標決定式得自陳量表	因素分析自陳量表	理論依據的自陳量表
孟氏行為困擾調查表: 作為個別諮商&團體討論用。 **少年人格測驗:** 輔導學生的參考。	利用效標組與正常人(或稱控制組)在人格量表題目的作答反應的統計差異顯著性檢定，以選擇最能有效區別控制組與效標組的題目而形成量表。	針對受試者施以大量的人格測驗的試題，透過因素分析法認定依些共同因素，而將題目依這些因素結合再一起。	編制者依人格理論詳細界定各項人格特質的意義，然後編擬可以代表這些特質的題目。

- **評定量表**(personality rating scale)
  - **性質**

    人格評定量表係將所測量的人格特質，列舉許多有關的題目或問句，由評定者根據他對被評定者(受試者)的多方面觀察結果，加以評定。
  - **偏誤**

月暈效果	中庸傾向的偏誤	慈悲偏誤&嚴苛偏誤
Halo effect	Error of central tendency	Leniency error & severity error

- **投射測驗**(projective personality test)

聯想法	完成法
羅夏克墨漬測驗(Rorschach Inkblot Test) 文字聯想測驗(Galton)(word association test) ---自由聯想測驗	係提供一些不完全的刺激，由受試者去談補或完成。ex:**語句完成測驗**(sentence completion) **羅德**(J. B. Rotter)**未完成語句**(Incomplete Sentence Blank) **謝克句子完成測驗**(Sacks Sentence Completion Test)
編造法	
要受試者根據所得到的圖畫編造一套還含有過去、未來和現在等發展過程的故事，從而衡鑑受試者人格特質的方法。 **主題統覺測驗**(Thematic Apperception Test, TAT)---莫瑞(H. A. Murray) & 摩爾根(C. D. Morgem) **羅氏逆境圖畫測驗**(挫折圖形研究)(Rosenzweig Picture-Frustration Study, P-F Study)	
表現法	**選擇法**
畫人測驗(Draw-a-Person Test, DAP)	宋狄測驗(Szondi test) 卡賽爾(K. N. Cassel)與康恩(T. C. Kahn)的團體人格投射測驗(Group Personality Projective Test, GPPI)

- **情境測驗**(situational test)
  情境測驗是預先布置一種真實情境，觀察受氏置身該情境下所表現的實際行為，然後對其人格特質加以評鑑的方法。
  又可分: 日常情境測驗、情境壓力測驗、無領袖團體討論
- **語意分析(區別)技術**(semantic differential techiques)
  由奧斯古(Osgood, 1957)所創，目的是在測量與比較概念的意義。
- **Q 技術(Q 排列法)**
  由史蒂芬生(Stephensonm 1953)所創用，可以用來評量態度、興趣、自我觀念和其他情感變相的一種測量方法。設計行為的特質。
- **態度量表**
  可分為認知、情感和行為的。方法兩種: 一為觀察法；二為自陳量表。

興趣測驗

興趣與性向

興趣是指個人對某些事物或活動有所喜好且主動接觸、參與的積極心理傾向，興趣是學習的動力，也是學習的結果。

性向是先天的能力，較穩定；**興趣**是一種動機，易受環境影響。

職業興趣測驗的發展趨勢

強調自我探索	強調拓展生涯選項	強調性別平等待遇

興趣量表簡介

何倫(Holland)興趣量表			庫德興趣量表	史東興趣量表
實際型	研究型	藝術型	Kuder Preference Records	Strong Interest Inventory
社會型	企業型	傳統型	選修學科選組、選系、職業選擇的參考	普通職業主題、基本興趣量表、職業量表、施測指數和特別量表。

適性測驗

- **組成要素:** 項目反應模式:單參數、雙參數或三參數

題庫	測試的起點	試題選擇方法	評分方法	測試終止表準

- **評論**

優點	缺點
(1) 適應考生的能力 (2) 考試的實行可以標準化 (3) 能夠獲得傳統測驗不能提供的資訊 (4) 能提供及時的回饋 (5) 能使用新題型	(1) 必須按續作答，不能跨越題目 (2) 同時進行測驗的考生數量將取決於能提供電腦有多少 (3) 題目形式的多樣化能否實現取決於使用電腦的技術和能力 (4) 分數解釋的困難

第四節 班級測驗的編製

一、準備測驗編製計畫

- 準備測驗編製計畫
- 編擬測驗試題
- 試題與測驗的審查
- 試題與測驗的分析
- 測驗的編輯

確立測驗目的和目標

安置性測驗	Placement test	評量學生的起點行為，以便學生安置在適當的教學計畫中。
形成性測驗	Formative test	了解教學過程中學生的學習進步情形，隨時提供回饋給教師和學生。
診斷性測驗	Diagnostic test	診斷學生的學習困難，以做為補救教學的依據。
總結性測驗	Summative test	評量學生學結束時的成就表現，以確定學生是否精熟，而達成教學預定目標。

設計雙向細目表

雙向細目表(two-way specification table) --- 教學目標 & 教材內容

選定測驗的題型

選擇型試題(客觀測驗)—selection-type items	**補充型試題(論文測驗)**—supply-type items
是非題、選擇題、配合題、填充題、解釋性試題、重組題	簡答題、限制反應題、申論題

選擇型試題

是非題—艾柏(R. L. Ebel, 1971)

編製原則	優點與限制
1. 文字的敘述力求簡要，避免冗長複雜 2. 避免使用含糊不確定的數量用語 3. 避免使用否定的敘述，尤其是雙重的否定 4. 同一題中的敘述避免包含兩個概念，除非測量因果關係的題目 5. 在因果關係的題型中，結果的敘述必須是對的，而原因的敘述可對可錯 6. 對與錯的題數宜大致相等，且隨便排列 7. 對的題目與錯的題目在敘述的長度上應接近相同	**優點** 1. 涵蓋教材內容的範圍比其他客觀試題大。 2. 容易命題、效率高、迅速客觀 3. 辨認因果關係的能力 4. 測量有誤解的信念或迷信 5. 有誤解的信念或迷信 **限制** 1. 容易猜測、信度低容易誤解 2. 容易形成自動答對、答錯

- **選擇題**

是由題幹(stem)與選項(option)所構成。其中有誘答選項(distracter)，練習「再認」的能力，需要某種以上的精熟程度，選出最佳選項。

編製原則	優點與限制
1. 題幹的敘述須能清楚顯示出題意，避免冗長 2. 題幹力求完整 3. 少用否定語句 4. 選項力求一致，3-5 之間 5. 所有誘答項目應具似真性 6. 少用「以上皆非」&「以上皆是」 7. 正確答案出現在各選項平均	**優點** 1. 容易配分，信度較高，猜答率較小 2. 難度容易調整，可避免反映心向 **限制** 1. 命題困難、容易偏重記憶 2. 無法綜合組織技術面 3. 僅限

- **配合題**

由前題題目(premise)與反應項目(response)組合而成。

編製原則	優點與限制
1. 僅用同質性的材料 2. 問題項目與反應項目不宜相等 3. 問題 5-8，反應 over2-3 4. 反應項目宜有系統的 5. 作答方法要寫清楚 6. 題目力求簡短，問左反右 7. 印在同一頁上	**優點** 1. 精簡試提供多樣化的反應項目 2. 分享同類反應項目的選擇題來改編 3. 計分容易、客觀且可靠 **限制** 1. 使用無關線索較多 2. 簡單記憶或機械記憶的事實知識 3. 編制題目困難

- **填充題**(簡答題)

編製原則	優點與限制	
1. 一試試題一個答案，越精簡越好 2. 採直接問答句 3. 空白必須是重要概念 4. 空格不應太多 5. 放置末端	**優點** 1. 可以減少學生採測 2. 容易編擬 3. 適合測量計算題	**限制** 1. 需要人工計分，評分難 2. 以事實知識為主

- **解釋性試題**(interpretive exercise)

常以題組形式出現。--- (1)閱讀式 (2)圖表式

編製原則	優點與限制
1. 根據教學目標挑選文章，須符合程度 2. 要能測到複雜的效果 3. 文章必須新穎、簡短，具可讀性 4. 試題數量及導讀文章成比例 5. 若用解決式題目，問題分類必須獨力相斥	**優點** 1. 可測到複雜效果 2. 可變化不同題型、計分容易 **限制** 1. 文章難找、能力問題、容易提供線索

	2. 無法測量學生創造、組織及表達

➢ **重組題**

編製原則	優點與限制
1. 項目 3-7、個項目隨機排列加以編號 2. 只有一種正確排列方式	**優點** 作答簡答，猜題率低、編擬容易 **限制** 除語文、歷史、生物外其他科目南出題、計分爭議多

補充型試題

➢ **類型**

與客觀式測驗相對的就是勇許學生自由反應的論文題，亦即建構反應式(constructed response)試題，可分為兩類

- 完全自由，沒有限制的申論題(extended response type)
- 局部限制，有限制的申論題(restricted response type)

高度受限制的反應 ←----------→ 高度自由的反應				
試題類型	簡答題、計算題	問答題、解釋名詞、數學應用題	申論題	建構式作業、創造思考作業、寫作
主要認知層次	知識、計算技巧	知識、理解、應用	分析、綜合、評鑑	應用、綜合、評鑑

➢ **申論題編製原則 & 優點與限制**

編製原則	優點與限制
1. 較複雜較高層次的評量 2. 明確界定測量的行為，議論範圍要適當 3. 問題要說清楚，以免各有不同見解 4. 不要採用可以自由選題的作答方式 5. 給予充分的作答時間，提示每題時限 6. 多提短答，不要長答方式	**優點** 1.能夠測量複雜效果 2.能明顯影響學生的學習方法 3.比較節省編擬試題時間 **限制** 1. 試題抽樣少、內容效度低 2. 計分費時、主觀且信度低、有語文因素影響

➢ 申論題的評分原則

- 妥善預擬一份評分要點，作為依據
- 使用最適當的評分方法，分數法(point method)、等級法(rating method)
- 依據所預期的學習結果來評分，避免受到無觀因素影響
- 一次只評閱一份試題
- 一再同一時間評完所有考卷
- 兩位以上的評分者較佳、匿名方式評分

二、編擬測驗試題

1.試題取材均勻，並求最大的內容效度	5.試題必須以概念為中心重新組織，避免直接抄自課文或原來教材
2.試題應重視重要概念或原理原則的理解與應用，而非零碎片段知識的記憶	6.教材前後有關聯的內容一統合再一起命題，以測量融會貫通的能力
3.第一次草擬的題數最好超過最後定稿所需的題數	7.試題的敘述應力求簡明扼要，題義明確
4.各試題一彼此獨立，互不干涉，並避免含有暗示性答案	8.試題必須有公認的正確答案，避免有爭論的答案存在

三、試題與測驗的審查

進行審查(review)可分為兩種

- 邏輯的審查(logical review)
- 實證的審查(empirical review)

邏輯的審查

主要測量是否有一致性(consistency)與適當性(adequacy)

一致性(consistency)	適當性(adequacy)
1. 試題是否代表所要測的目標 2. 試題與教學目標是否一致 3. 試題與教學的呈現是否一致	檢查重點在於查驗試題的格ㄟ、問題陳述的品質、以及其他可能的影響因素是否能夠適切的反印出試題所要測量的行為目標。
Rocinelli & Hambleton(1977)提出一種計算試題與目標間是否匹配的指標，稱作「試題與目標一致性」(item-objective consistency, IOC)指標。 **值域:** -1.0(非一致) ~ +1.0(一致)	檢查試題內容、題署、範圍是否遵照雙向細目表。

實證的審查

試題分析(item analysis)

- 難度指標(difficulty index)
- 鑑別度指標(discrimination index)
- 然當代的試題反應理論(item response theory)認為還有猜測度指標(guessing index)屬一種不辨測量值(invariant measure)

教學敏感度分析(instructional effectiveness)

- 最常用的是**前後差異指標**(pre-to –post difference index, PPDI)(Haladyna & Roid, 1981)
- PPDI = (後測難度指標) – (前側的難度指標)
- **值域:** -1.0 ~ +1.0 (正常 PPDI 指標介於 0.10~0.60)
- **作用:** 1.保留　2.刪除　3.修改該試題

四、試題與測驗的分析

試題分析部分(項目分析)	測驗分析部分
1. 難度指標(difficult index)的分析 2. 鑑別度指標(discrimination index)的分析 3. 誘答力(distraction)的分析 4. 注意指標(caution index)的分析	1. 信度指標(reliability index)的分析 2. 效度指標(validity)的分析 3. 差異指標(disparity index)的分析

試題分析

1.提供回饋給學生	2.作為教師實施補救教學的依據	3.作為修改課程建議的憑據
4.增進教師命題的技巧	5.增進測驗題庫運用的效能	

常模參照測驗的試題分析

➢ 試題分析的步驟

- 按照總分高低排列，取前後 25-33%
- 分別計算出高分組兩滴分組在每一事題的答對人數及百分比
- **難度指標**

$$P_i = \frac{P_{iH} + P_{iL}}{2}$$

- **鑑別度指標**

$D_i = P_{iH} - P_{iL}$

- 若是選擇題，檢視每一選項可得出誘答力。

➢ 難度指標的分析

- **答對百分比法**(number correct ratio)---item difficult index

$$P_i = \frac{R_i}{N} \times 100\%$$

$$P = \frac{P_H + P_L}{2}$$

- **范氏試題分析表法**(Fan's item analysis table)(Educational Testing Service, ETS)

$\Delta = 13 + 4Z$

Z: 常態分配量尺上的標準分數　　13: 代表轉換公式的平均數

4: 代表轉換公式的標準差　　值域: 1~25 平均難度 13

Δ 越大，愈困難；Δ 越小，愈簡單。

- **試題難度與測驗分數的分配**

題目簡單，P 值大，峰態呈現負偏態。

題目困難，P 值小，峰態呈現正偏態。

➢ 鑑別度指標的分析

分兩類: 內部一致性(internal consistency)法及外在效度(external validity)法

- **內部一致性分析法**

瞭解各試題的功能是否和整個測驗的功能相同。

$D = P_H - P_L$

D: 鑑別力指數　P_H: 高分組答對百分比　P_L: 低分組答對百分比

值域: -1.0～+1.0 (數值大鑑別力高，數值小鑑別力低)

- **外在效度分析法**

學生在每個試題上的反應與其在校標上的表現所具有的相關情形，也可以做為試題鑑別度的編制。

點二系列相關法	二系列相關法	ϕ 相關法
數值高，鑑別度高	常態分配、分為答對答錯二種	試題與校標均是二分項的情況

- **誘答力的分析**(distraction analysis)

校標參照測驗的試題分析

- 難度指標分析(不須做難度分析，通過率即難度)
- 鑑別度指標分析
- 誘答力分析

試題的選擇

- 難度與鑑別度的關係(P=0.5；D=0.5)
- 挑選優良試題的標準

要項	名稱	值域	保留原則
邏輯審查	試題與目標一致性(IOC)	-1.0～+1.0	愈接近 1 愈一致
實證審查	教學敏感前後差指標(PPDI)	-1.0～+1.0	正常值 0.10~0.60
試題難度分析	難度指標(指數)(P)	0～1	接近 0 困難
試題鑑別度分析	鑑別度指標(指數)(D)	-1.0～+1.0	指數高，鑑別力大 0.40 以上為優良

當代理論的分析

- **IRT 理論**
  - **IRT 理論及優點**

優點	適用條件
1. 試題難易不因樣本有異而改變 2. 學生程度不因測驗難度而不同 3. 測量標準誤不因學生不同 4. 可應用多項測量問題	1. 測驗品質須先建立 2. 能先預試 3. 要有足夠的樣本數可供預試 4. 資料與模式須相符、分析困難

  - **IRT 的假設**

單維性	局部獨立性	要適合模式的要求	答題的時間不受限制

  - **IRT 的應用**

測驗的等化	編製測驗	建立題庫	適性測驗	試題偏向的檢驗

- **S-P 表分析**(student-problem chart analysis theory, S-P)
  - 作答反應資料的重要性
  - S-P 表分析的理論要義—無母數檢定、同質性係數、試題注意係數、學生注意係數
  - S-P 表的編制與涵義—S 曲線以左、P 曲線以上；S 曲線以右、P 曲線以下，稱為「完美量尺」。
  - 差異係數(disparity index)
  - 注意係數(caution index)

五、測驗的編輯

測驗的題數

測驗的目的	試題的類型	信度的高低	學生的年齡	學生的能力	作答的實現

試題的難度---難度與範圍有關

- **試題的排列**

根據試題難度來排列	根據教學目標或測量能力來排列	根據試題類型來排列	根據教材內容來排列

- **編制測驗指導語**

六、測驗的實施與計分

- **評量工具的選擇**

信度	效度	區辨力(discriminatory power)	公平性	實用性

- **影響測驗使用的因素**

<table>
<tr><td>1.試題品質的好壞</td><td>2.施測程序和情境</td><td>3.主測者因素</td></tr>
<tr><td rowspan="3">4.反應心向(response set)</td><td colspan="2">速度—正確的反應心向</td></tr>
<tr><td colspan="2">賭徒式的反應心向</td></tr>
<tr><td colspan="2">與命題有關的反應心向</td></tr>
<tr><td>5.作弊</td><td>6.動機與焦慮</td><td>7.應試技巧</td></tr>
<tr><td>8.教導與學習</td><td>9.月暈效應(halo effect)</td><td>10.測驗偏差</td></tr>
</table>

第五節 教學評量

一．教學評量的概念

- **意義**

  美國教學量學家吉勒(R. J. Kibler, 1978)在《教學目標與評量》一書提到了「教學基本模式」(The General Model of Instruction, GMI)。

- **內涵**
  - 教師教學效率的評量
  - 學生學習成就的評量—學期前、中、後評量
  - 課程設計與實施的評量
- **功能**
  - 可了解起點能力
  - 診斷學生學習困難與障礙
  - 了解潛能與學習成就
  - 估量教師教學的效率，作為教師改進教材、教法的參考
  - 學生自我評量的結果

二．教學評量的種類

依評量的時機與作用區分	形成性評量(formative evaluation)	過程評量
	總結性評量(summative evaluation)	單元課程評量
依評量的功能區分	安置性評量(placement evaluation)	預備性評量
	診斷性評量(diagnostic evaluation)	未診斷出問題點
依評量結果的解釋方式區分	常模參照評量 (norm-reference evaluation)	
	校標(標準)參照評量 (criterion-reference evaluation)	
依評量所使用的工具和形式區分	紙筆測驗(paper-and-pencil test)	
	變通性(另類)評量(alternative assessment)	

	常模參照評量	**校標參照評量**
主要的用途	綜合性評量	精熟性的測量
主要的重點	測量成就的個別差異	敘述學生能做的工作
結果的解釋	和別人的成就表現比較	和具體明確的成就領域比較
涵蓋的內容	涵蓋廣大的成就領域	集中在有限的學習內容
測驗計畫的性質	使用雙向細目表	使用詳細的教材領域細目表
試題選擇的方法	選擇最能區分個別差異的題目(分數的變異性最大)。刪除容易的試題。	包含所有能適當敘述表現的題目。會改變題目的難度，或刪除容易題目，以提高分數的變異性。
成就的標準	依據在團中的相對地位來決定成就水準。	依據絕對的標準來決定成就的水準

三、教學評量的基本原則

- 依據教學目標
- 宜顧及學生的全面發展
- 應採多元、多次評量
- 強調反應歷程
- 妥善應用評量結果

四、情意領域的評量

- **忽視情意目標的原因**

教學上的困難	評量技術上的困難	擔心造成負面效果
目標的含糊性與長期性 擔心造成價值觀的灌輸 社會對多元價值觀的包容與尊重	學生誠實回答的意願 學生回答問題的能力 資料蒐集過程過程無法客觀 外在情境影響大	侵犯隱私權 逼使學生更加虛偽

- **情意領域的評量方法**

觀察法	社會計量法	自陳量表法	投射技術
Observation techniques	Sociometric techniques	Self-report inventory	Projective technique
軼事記錄法 評定量表法 項目檢核表法	社會計量矩陣 社會關係圖 猜是誰技術	自傳 人格量表 興趣量表 態度量表	聯想技術 完成技術 編造技術 表現技術

- **與情意發展相關的學習環境**
  - **班級氣氛**

民主與尊重	自由與開放	關懷與親密	價值與規範	目標與績效

  - **人文取向學習環境之評量**(人本主義)

教師少支配學生	教師少說話	教師少問有正確答案問題	評量是師生共同實施	學生學習動機高

- **多元化教學評量**
  - **評量的趨勢**

靜態評量(static assessments)	>>>	動態評量(dynamic assessments)
機構化評量(institutional assessments)	>>>	個人化評量(individual assessments)
單一評量(single assessments)	>>>	多元評量(multiple assessments)
虛假評量(squrious assessments)	>>>	真實評量(authentic assessments)

  - **多元化評量的意涵**
  - **教學評量的發展趨勢**

教學與評量統合化、適性化	評量專業化、目標化	評量方式多元化、彈性化
評量內容生活化、多樣化	評量人員多元化、互動式	結果解釋人性化、增強化

多元評量的方法

➢ 檔案評量(portfolio assessment)

檔案評量的特性

容易達到教學目標	重視學生個別差異	要求組織與統整
評量多元化	學生餐與對自己的評量	強調找出學生的優點而非弱點

適用時機

- ✓ 班級數少，師生頻率高
- ✓ 教師兼有共識，願意配合做長期的觀察和紀錄
- ✓ 學生間能力差異大，有必要進行個別化的教學與評量

檔案評量的內容

檔案的類型	檔案評量的內涵(方法)
成果檔案 過程檔案 評量檔案	行為觀察評量表(評量學習過程) 作品及測驗結果(評量學習結果) 晤談紀錄(評量情義成果與學習困難診斷) 角色扮演評量表(評量人際關係的能力) 模擬演練評量表(評量處理事情的能力) 個人重要事件摘要(軼事紀錄)

評量規準

達成學習目標的程度	展現個人特色的程度	與他人合作的程度	運用資源了能力

檔案評量的優缺點

優點	缺點
1. 強調自我成長，培養主動積極 2. 呈現多元資料，激發創意 3. 兼顧認知、技能、情意等體評量 4. 獲得教真實的評量學習結果 5. 兼顧歷程與結果的評量 6. 動態歷程激發學習興趣 7. 增強學生溝通表達與組織能力	1. 檢閱學生作品費時 2. 製作需多經費 3. 若無協助對象，難以發揮優點 4. 評分者間的一致性偏低 5. 結果易受月暈效應影響，降低評量效度

以學生為主的親師會談

會談目的	實施步驟	會談重點

➢ **實作評量**(performance assessment)

■ **實作評量的性質**

目的	客觀測驗	論文測驗	口試	實作評量
目標	知識樣本，具有最大的有效性和信度值	評量思考技能或知識結構的精熟程度	評量教學中的知識	評量將知識和理解轉換成行動的能力
學生的反應	閱讀、評鑑、選擇	組織、寫作	口頭回答	計畫、建構和傳送原始的反應
優點	有效—能同一時間內進行多個測驗試題的施測	可以評量較複雜的認知成果	可以連結評量與教學	提供實作技能充分的證據
對學習的影響	過度強調記憶；如果妥善編制，亦可測量到思考技能	激勵思考和寫作技能的發展	刺激學生餐與教學，提供教師有關教學成效的立即回饋	強調在相關的問題背景情境下，使用現成技能和知識

■ **實作評量的目的**

彌補傳統紙筆測驗的不足	對學生的學習成就作更正確的推論	對教師的教學活動產生正向的引導

■ **實作評量的特徵**

實作的表現	實作評量的類型
1.實作的表現 2.真實性與直接性 3.問題情境的含糊性與表現的彈性 4.兼重評量的結果與歷程 5.多向度的評分系統 6.評分以人為判斷為主	1.教學目標的特性 2.課程的教學程序 3.客觀條件的限制

✓ **Gronlund**(1993)依據情境的真實程度，將教學情境常用的實作評量分成下列五種類型

紙筆式實作測驗	Paper-and-pencil performance
實務辨認測驗	Identification test
結構式實作測驗	Structured performance test
模擬表現	Simulated performance
工作樣本	Work sample

模擬表現

學習負荷量(overload)	學習遷移(transfer)
學習動機(affect)	製作成本(cost)

- ✓ **Linn&Gronlund**(1995)
  - 限制反應實作作業(restricted-response performance task)
  - 擴展反應實作作業(extended-response performance task)

■ 實作評量的實施步驟

- ✓ 確立實作評量為目的
- ✓ 確認實作評量的標準
- ✓ 提供適當的表現情境
- ✓ 選擇計分和評定方法

檢核表(checklists)	評定量表(rating scales)
系統的觀察和軼事紀錄(systematic observation and anecdotal records)	
作品量表(product scales)	檔案評量(portfolios assessment)
口語評量(oral assessment)	遊戲化評量

■ 實作評量的評分與限制

- ✓ **評分**

整體性評定法(global rating)	**分析式評定法**(analytic rating)
作品等第量表排列法 等第排列法 心像比較法	檢核表 評定量表 評分規程

- ✓ **限制**

偏見(bias)	月暈效應(halo effect)	評量次數過少

- ✓ **改進方法**
  - 可以被觀察到、測量到、與進行數字量化的術語，說明預期結果
  - 擬定清楚、明確的且適合學生程度的標準
  - 標準個數控制在 10-15 以內
  - 選擇足夠學生表現的真實環境
  - 使用結構化施測情境
  - 紀錄要客觀、多觀察

■ 實作評量的優缺點

優點	**缺點**
1. 能同時評量認知與技能方面的教學目標 2. 能提供技能學習方面的診斷資料 3. 接近現實生活、增進學習遷移 4. 直接測量，排除語文能力的干擾	1. 實施上費人力，時間及金錢，不合經濟效益 2. 測驗情境困難 3. 計分不客觀 4. 信度與概化 5. 對容易焦慮的學生不利

➢ **真實評量**(authentic assessment)

■ **意涵**

測試結構	設計情境	分級和計分

■ **特性**

改變學生的角色	改變教師的角色	增進家長的積極角色

■ **評量真實性的準則**

強調評量的內涵與生活非常貼近。

➢ **動作評量**(dynamic assessment)

■ **意涵**

✓ 傳統評量以心裡劑量理論為基礎，提供學生單一時間點表現或成就的相對地位訊息，故又稱「靜態評量」(tatic assessment)

✓ 由 Feuerstain 在 1979 年提出，內涵

- 了解受試者動態認知歷程與確認認知能力的變化情形，著重認知及學習歷程的改變
- 著重評量者與受試者的互動關係，強調評量與教學結合，教師運用「前測—教學介入—後測」的主動介入模式，經由充分溝通互動歷程，持續觀察。

■ **理論基礎**

Vygotsky 的社會發展認知論為基礎。社會文化層面來探討學習與發展的交互關係，提出社會中介(S-O-R)、內化、近測發展區等。

S-O-R---------「刺激—個體—反應」

S-H-O-H-S---「刺激—人際社會互動—個體—人際社會互動—反應」

■ **特性**

兼重學習結果與學習歷程	兼重回溯性評量與前瞻性評量	兼重鑑定、診斷與處方
著重認知能力的可塑性	著重師生雙向溝通的互動關係	融合教學與評量

■ **類型**

Feuerstein	Campione & Brown	Burns, Vye & Bransford
學習潛能評量設計	**漸進提示評量**	**連續評量**
強調中介學習是認知條件，以診斷學生認知功能缺陷，評量學生對教學反應為目的。 **「前測—中介—後測」**	認為動態評量不僅評量過去已有的經驗、技能或知識，更應評量成長、改變和學習預備度。 「由一般、抽象逐漸到特定、具體」 **「前測—學習(訓練)—遷移—後測」**	認為中介學習乃增進認知發展的重要條件，以檢視不同教學介入效果，確認有效介入成分為評量目的。 **「前測—訓練—再測—訓練—後測」**

評析

優點	缺點
1. 比較不會低估文化不利、身心不利的學生 2. 比較能了解學生如何表現的連續性學習歷程，較能確認學生思考歷程與解決策略的缺失 3. 較能顧及學生的個別差異，尊重學生思考 4. 較能避免非認知因素對教學與評量的干擾 5. 較能強化學生正向自我觀念	1. 不容易執行、個別化評量成本高 2. 前測訊息未充分使用 3. 教學介入內容缺乏理論依據 4. 信度與效度有待加強

評量結果的報告

學期成績報告的功能

回饋學生	提供學生申覆機會	與家長聯繫	分析學生錯誤概念	改進命題技巧

評定成績時的比較標準

相對比較(解釋)法		絕對比較(解釋)法		自我比較法	
係以全班學生成績的平均數為基準點，利用離均差的尺度，對成績在平均數以上的學生憑以較高的等級，對成績低於平均數者評以較低的等級。		係拿個人的表現去和事先設定的、理想的標準做比較，至於同儕表現的優劣，對於個人的成績毫無影響。		係基於自我求進步的原理，將某一學生前後幾次的測驗成績相比較。	
優點	**缺點**	**優點**	**缺點**	**優點**	**缺點**
可以區別學生學習的成就及努力的程度，並避免因試題難易所造成績偏高或偏低的現象。	忽略了學生的個別差異，評量未必公平。	有固定的標準，易於達成品質管理。 學生有努力的目標，只要努力即可。	標準訂定不易。 忽略個別差異。	適合個性原理，顧及個性差異。 可做教育準段的運用。 容易看出學生進步情形。	無法和某種參照標準，無法看出一個人的優劣。 無法做學生間的比較。 兩次的考試難度會不一樣。

成績評量結果的表示

百分制		等第制	
傳統計分方法，標準 60 or 70		利用統計學的常態分配原理 ABCDEF	
優點	**缺點**	**優點**	**缺點**
通俗易懂 單一數字，作業方便	難度不同，意義不同 得分不同，可能因誤差所致，無真正的差別	成績評量採等第制可趨勢學校常態編班，教學正常化。	若非常態分班會有前後段班成績不公平現象。

解釋測驗結果的原則

- 測得分數必須要有意義、需考量學生身心狀況及其他因素。
- 要用「一段分數」不是用「特定的數值」、須以其他的證據為佐證。

- **教育測驗的運用**
  - 篩選性評量與機構效率
  - 安置性評量與因材施教
  - 診斷性評量與補救處理
  - 形成性評量與精熟學習
  - 總結性評量與成績考核
  - **常模參照式評量與相對地位**

簡單常模表	多組別共用的常模表	多種分測驗共用的常模表
Simple norms tables	Multiple-group norms tables	Multiple-score norms tables
多種衍生分數共用常模表	簡略式常模表	濃縮是常模表
	Abbreviated norms tables	Condensed norms tables

  - **標準參照式評量與能力證明**

精熟學習	職業證照考試	基本學力測驗

  - 動態評量與學習潛能

- **國中基本學力測驗**
  - **緣起及理念**----85 年
  - **性質**---87 年推動
  - **測驗內涵**
  - **特色**

標準化	可比較	一年多試、一試多用	能力導向	對教學有良性影響的

  - **量尺分數**
  - **PR**

七、國內外評量計畫

- **國際學生能力評量計畫**(PISA)---經由 OECD 籌辦
  - **目的**

    針對十五歲學生，生活知能的學習成效提供跨國際的比較，以及各國教育效能的分析，並由此界定國民素質的內涵。
  - **內涵**

    PISA 是一項以年齡為導向，調查研究，主要在評估接近完成基本教育的十的五歲學生，能否接受社會的挑戰。
  - **程序**

    每三年一次，15 歲學生包含公私立學校。
- **美國國家教育發展評量**(NAEP)
- **促進國際閱讀素養研究**(PIRLS)
- **台灣學生學習成就評量資料庫**(TASA)